Robust Face Alignment Under Occlusion via Regional Predictive Power Estimation

Heng Yang, *Student Member, IEEE*, Xuming He, *Member, IEEE*, Xuhui Jia, *Student Member, IEEE*, and Ioannis Patras, *Senior Member, IEEE*

***Abstract*— Face alignment has been well studied in recent years, however, when a face alignment model is applied on facial images with heavy partial occlusion, the performance deteriorates significantly. In this paper, instead of training an occlusion-aware model with visibility annotation, we address this issue via a model adaptation scheme that uses the result of a local regression forest (RF) voting method. In the proposed scheme, the consistency of the votes of the local RF in each of several oversegmented regions is used to determine the reliability of predicting the location of the facial landmarks. The latter is what we call regional predictive power (RPP). Subsequently, we adapt a holistic voting method (cascaded pose regression based on random ferns) by putting weights on the votes of each fern according to the RPP of the regions used in the fern tests. The proposed method shows superior performance over existing face alignment models in the most challenging data sets (COFW and 300-W). Moreover, it can also estimate with high accuracy (72.4% overlap ratio) which image areas belong to the face or nonface objects, on the heavily occluded images of the COFW data set, without explicit occlusion modeling.**

***Index Terms*— Face alignment, occlusion, random forest, cascaded pose regression, model adaptation.**

I. Introduction

FACE alignment, or in other words the localization of a set of facial landmarks, such as the center of the pupils or the tip of the nose in a face image, is a well studied topic in the computer vision literature. The interest in automatic localization of the landmarks lies in many important applications such as face recognition, facial animation and facial expression understanding. In recent years, significant progress has been made in this task and several works have reported very good results on datasets collected in the wild [8], [34]. Nevertheless, most of the face image datasets do not have significant occlusions, for example, the widely used Labelled Facial Parts in the Wild (LFPW) dataset [6] has an average of 2% occlusion. Face images in real scene, such as the ones in the recently created COFW dataset [7], are often more challenging. These methods performed significantly worse when applied on such images with heavy occlusion since their models cannot handle missing features due to occlusion. Despite the fact that face images in real world are frequently occluded by objects like sunglasses, hair, hands, scarf and other unpredictable items, as shown in Fig. 1, very few works have studied face alignment under occlusion explicitly.

Tackling the occlusion problem explicitly is difficult mainly due to two reasons. First, compared to the intra-category shape variation of face, the occluders[1] are much more diverse in appearance and shape. They can appear on the face in almost unpredictable arbitrary position with various sizes. Second, it is a chicken and egg problem since that occluders should not participate in the alignment but it is difficult to tell whether a landmark is occluded unless the correct alignment is known [26]. Therefore, most of the existing works only considered the occlusion status of individual landmarks and treated the occlusion landmark as unstructured sources of noise. In addition, they require the annotation of occlusion during training, either annotated manually [7] or synthesized artificially [18]. These approaches show some success but have a series of drawbacks:

- Treating the occlusion status of individual landmark independently ignores a key aspect that the occluders are often other objects or surfaces and hence often appear in continuous regions instead of an isolated pixel.
- The randomly synthesized occlusion patterns are not realistic enough to describe the occlusion diversity in real scenes. To collect face images with occlusions and to annotate their occlusion status is expensive, especially when a large number of such images are demanded for model training.
- The occlusion detection at pixel level limits its practical application in face analysis since features are usually extracted from a region rather than an individual pixel.

The method presented in this paper aims to deal with face alignment under occlusion and overcome the above mentioned drawbacks. An overview of our method is shown in Fig. 1. Given a face image, our method starts from a detected face and

Manuscript received July 11, 2014; revised October 28, 2014, January 15, 2015, and March 24, 2015; accepted March 30, 2015. Date of publication April 10, 2015; date of current version April 29, 2015. This work was supported by the European Integrating Project REVERIE under Grant FP-287723. The work of H. Yang was supported by a China Scholar Council Scholarship. The work of X. He was supported in part by the Australian Government through the Department of Communications and in part by the Australian Research Council through the ICT Centre of Excellence Program. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Stefano Tubaro.

H. Yang and I. Patras are with the Queen Mary University of London, London E1 4NS, U.K. (e-mail: heng.yang@qmul.ac.uk; i.patras@qmul.ac.uk).

X. He is with National ICT Australia (NICTA) and Australian National University, Canberra, ACT 0200, Australia (e-mail: xuming.he@nicta.com.au).

X. Jia is with The University of Hong Kong, Hong Kong (e-mail: xhjia@cs.hku.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2015.2421438

[1] In this paper the objects that occlude the face are called occluders and the visible face region is called face mask.

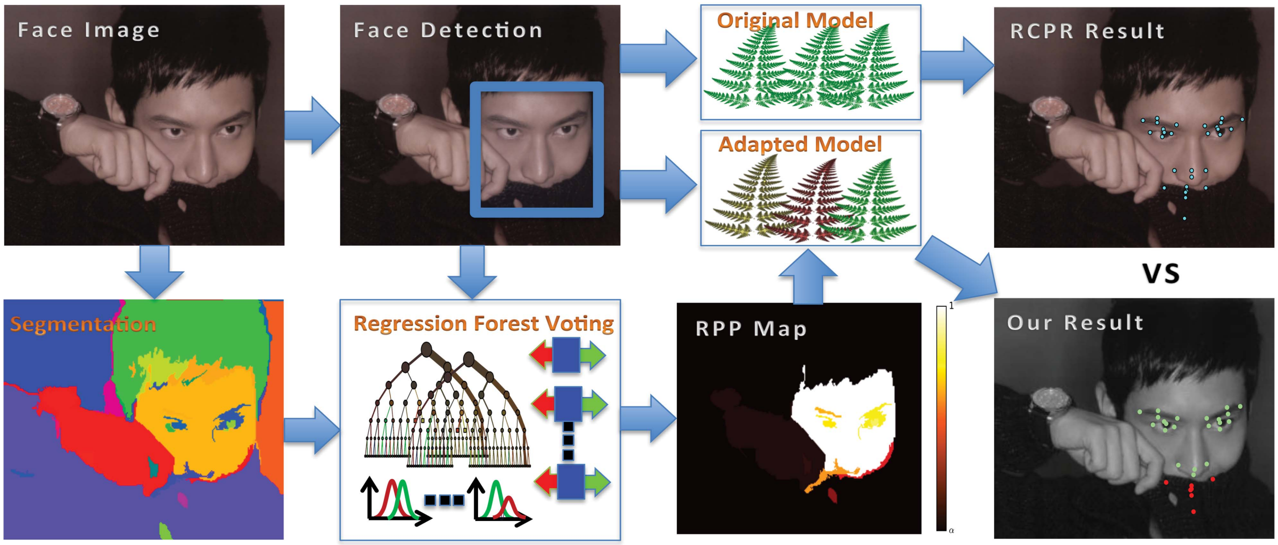


Fig. 1. Illustration of the pipeline of the proposed method. Given a test image, we first detect the face and apply segmentation by the graph-based approach in [16]. Based on the face bounding box information and the segmentation result, we employ the local patch based Regression Forest voting method for face alignment and obtain the Regional Predictive Power map with pixel probability from α to 1. We then adapt the state of the art face alignment model, (Robust Cascade Pose Regression (RCPR) is used as an example) by putting weights on different weak regressors. The final column shows the results from original RCPR (upper) and the adapted RCPR (lower). Our method is able to localize the landmarks more accurately (especially when occlusion is presented) and reason the occlusion labels of the landmarks (green = unoccluded, red = occluded).

employs an over-segmentation method to partition the image into non-overlapping regions. Then a local regression forest voting based facial feature detection approach is adapted to predict the power of each region affiliated to the face bounding box. We call this the Regional Predictive Power (RPP) and is essentially a measure of how useful information from a certain region can be for the task of face alignment. The output of this step is a dense RPP map that also indicates the probability of each region belonging to the face. This RPP map is then used along with the original face image for final face alignment using an adapted Cascaded Pose Regression methods. In summary, we make the following contributions in this work:

- We reason about the face mask (occlusion), represented by the RPP map, in an unsupervised manner, i.e., we do not use any occlusion annotation or synthesize occlusion patterns for model training but rely on the consistency of the local Regression Forest (RF) voting, that is the RF model is pre-trained for general facial feature detection. It follows a patch-based Hough voting scheme, such as [13], [38]. The occlusion prediction is at regional level and it holds two important properties that differentiate it from the previous works: first, it is dense, i.e., each pixel inside the face bounding box has a probability that indicates its confidence level of belonging to the face; second, it is structured, since the structure of both the face region and the occlusion pattern is naturally kept via the over-segmentation process.
- We adapt the recent face alignment model by taking the RPP map into account and make it more robust to partial occlusion. The core idea is that we use the occluded or erroneous features in the alignment process in a different way. The adaptation is made only during the testing stage, i.e., the model we are going to adapt is pre-trained and no additional annotation or re-training is required. We test this adaptation scheme on the state of the art face alignment approach, i.e., the Cascaded Pose Regression (CPR) and show clear improvement.
- We propose an initialization scheme that derived from the local RF detection, which improves the robustness to the face bounding box shift from training to testing stage.
- We extend the COFW dataset by manually annotating the face mask for each image, which can be used for evaluation of face mask prediction for further research.

We evaluate the proposed method on two most challenging datasets, namely, COFW dataset [7] and the 300-W benchmark dataset [27]. We show better or comparable results when comparing with the state of the art methods in the problem of face alignment in both datasets. Moreover we also show that we can estimate with high accuracy which image areas belong to the face and which not - on the heavily occluded images of the COFW dataset the overlap ratio is 72.4%.

The remainder of the this paper is organized as follows: In Section 2, we briefly review the existing facial feature detection techniques related to our work. In Section 3, we first describe the Regression Forest based Regional Predictive Power estimation scheme and then present how we use the RPP map to improve the robustness of the face alignment method to occlusion. In Section 4, we show the experimental results of our proposed method on different face alignment 'in the wild' databases. We close with concluding remarks in Section 5.

II. Related Work

Two different sources of information are typically used for face alignment: face appearance (i.e., texture) and shape information. Based on how the spatial shape information is used we categorize the methods into local-based deformable model methods and holistic pose regression methods. The methods in the former category usually rely on discriminative local detection and use explicit deformable shape models to regularize the local outputs. The methods in the latter category directly regress the pose (locations of a set of landmarks) in a holistic way. We first briefly review face alignment methods in these two categories, and then discuss the issue of partial occlusion.

Local based deformable models usually need to train a discriminative local detector for each facial landmark. Many classification and regression methods are utilized in this framework, e.g. the Support Vector Machines (SVM) in [6] and [23] and Support Vector Regression in [20]. Recently Regression Forests (RF) [10], [13], [36], [37] were also used where the location of facial point is estimated by accumulating *votes* from nearby regions. Smith *et al.* [29] also follows a voting scheme based on exemplar images retrieval. Different types of image features are used, e.g., Gabor feature [33], SIFT [6], HoG [46] and the multichannel correlation filter responses [17]. Although some methods make no use of the shape information [13], it is common to combine the local detection with shape models since only a few facial landmarks are very discriminative and typically there exist multiple candidates for the location of one landmark. This can be done by using a shape model to either restrict the search region (see [20]), or by correcting the estimates obtained during the local search. Typical shape models include the Constrained Local Model (CLM) [3], [5], [10], [12], [28], the tree-structured model [18], [41], [44], [46]. Other optimization search methods are also applied to search for the best combination of the multiple local candidates, e.g. the RANSAC [6], [29], Branch & Bound [1], graph matching [45] and regression forest votes fine-tuning [38], [39]. Local based methods struggle under occlusion since the local detector is intrinsically sensitive to noise. Also when the number of the face landmarks increases, their efficiency for both training and testing drops sharply since the local detection is carried out for each landmark separately.

Holistic pose regression methods regard the pose as a whole and often align the shape in an iterative or cascaded way. A typical method in this category is the Active Appearance Model (AAM) [9]. At each iteration of the AAM fitting, an update of the current model parameters is estimated via a simple linear regression method. In this framework, better optimizations are proposed in [28], [31], [32], and [34]. Noticeable progress in iterative holistic shape alignment has been made in recent years in the framework of Cascaded Pose Regression (CPR) for instance [8], [14], [15] and face sketch alignment [40]. Those methods directly learn a structural regression function to infer the whole facial shape (i.e., the location of the facial landmarks) from the image and explicitly minimize the alignment errors in the training data. The primitive random fern regressor at each iteration employs shape indexed features as input. Recent iterative approaches include the work by Xiong and De la Torre [34] based on SIFT features, convolutional neural networks [30], the incremental cascaded linear regression [4] and the Local Binary Feature learning based cascaded method [25]. Most of the iterative methods in this category depend on the initialization that derived from the face bounding box. When the face detector changes, the performance usually drops sharply. Current CPR based methods like [7], [8], [14] attempt to deal with this issue by initializing the method with several shapes and then by selecting the median value of the outputs. Burgos-Artizzu *et al.* [7] proposes a *smart restart* scheme to improve the robustness to random initialization.

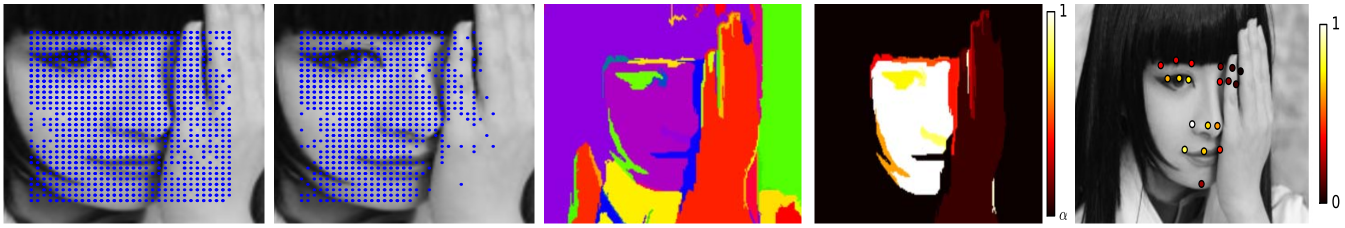
Partial Occlusion in face alignment has drawn very little attention. Local based methods have problem when heavy occlusion is presented because the local detector is inherently weak at dealing with occlusion. Then the global shape constraint usually leads to a local optimum. In contrast, the holistic methods can avoid the local optimum but they also struggle under occlusion since features that are extracted at occluded areas will directly affect the update of the whole pose at each iteration. It might result in a pose that is even far away from the true location. For instance, the AAM [9] is very difficult to deal with unseen images and occlusion. Only a few works have explicitly addressed the occlusion issue [18], [26], [35], [42], [43]. Those works focus on synthesized data or consider very limited number of occlusion patterns (sunglasses, scarf and hands). Those methods assume that only a small portion of the face image is occluded. However, in real scenarios, the occlusion patterns can be very diverse and are almost unpredictable. Burgos-Artizzu *et al.* [7] proposed an occlusion-centered approach that leveraged occlusion information to improve the robustness of the CPR method. It estimates the location of the landmark and, for each one an occlusion label, that is, whether it is visible or not. N visually different regressors are applied at each iteration. Each regressor is trained so that it uses features from only 1 out of 9 pre-defined image zones. During testing, the regressor outputs are weighed by weights that are inversely proportional to the occlusion prediction of the zone of each regressor. This method improves the CPR-based method [8], however, cannot deal with the large diversity of the occlusion patterns. In addition, all the above methods require additional occlusion annotations for training, that is expensive to obtain. Also they provide an occlusion label for each landmark, however, the occlusion often covers a region. In terms of predicting the importance on-line, similar idea was used in tracking [22].

III. METHOD

Our method consists of three main parts. In Section III-A we describe how we use the local Regression Forest voting scheme in order to predict the Regional Predictive Power (RPP) of regions that have resulted from an image (over) segmentation. In Section III-B we describe how the holistic Cascaded Pose Regression (CPR) face alignment model is adapted to a more difficult domain, i.e. the domain of occluded images, based on the estimated RPP. Finally, in Section III-C, we present the proposed initialization scheme.

A. Regional Predictive Power Estimation

It is challenging to directly model the face occlusion due to its unpredictable diversity in realistic conditions. However, the



(a) (b) (c) (d) (e)

Fig. 2. Regression Forest (RF) voting based Region Predictive Power (RPP) estimation. (a) shows the original votes distribution inside the face bounding box, similar dense for both the face region and occlusion region. (b) shows the distribution after the face center sieving [38]. As can be seen, many invalid votes from the non-face parts are effectively removed, which is a strong cue to predictive the RPP. (c) is the over-segmentation result. (d) shows the RPP map, i.e., the p_r in Eq. 2, calculated over each region of the segmentation. (e) is the detection result from the local RF model with the color varies according to the reliability of the landmark estimation, described in Section III-C.

occluders often occupy a continuous region and have different appearance than the face, or are separated from it by intensity edges. We use an over-segmentation and subsequently estimate a score that reflects the power/usefulness of each of the resulting regions in the face alignment task. The score is estimated by analysis of the votes of a local-based Random Forest algorithm, as shown in Fig. 2, and is closely related with the probability that the region in question belongs to the face.

We use the efficient graph based segmentation by Felzenszwalb and Huttenlocher [16], to get a set of regions, which ideally do not span multiple objects [2]. Let us denote with R_{SP} the set of superpixels and with $r \in R_{SP}$ a region in that set. The number of regions may vary from image to image. The RPP value of each region is generated in two steps as follows.

1) Sieving Votes in Regression Forest: We build the RPP prediction method based on the Regression Forest (RF) framework for face alignment, proposed in [13] and [38]. Image patch features that are extracted at several image locations cast votes for the localization of facial landmarks. As stated in [38], not all the votes from RF are reliable. Therefore, Yang and Patras [38] proposes to use a bank of sieves to remove unreliable votes based on the consistency by which they vote for the location of the face center.

More specifically, a set of patches is extracted from an input image I. Let us denote with V the resulting set of votes and by V_l the subset of the votes that are associated with the landmark l. Clearly, $V = V_1 \cup V_2 \cup ... \cup V_L$, where L is the number of landmarks detected by RF. Let us denote by V^r the set of votes that are associated with patches extracted within the region r. Each voting element $v = (\Delta_v, \omega_v, \Delta_v^o, \omega_v^o)$ consists of two types of voting information: one (Δ_v, ω_v) to a facial landmark and the other (Δ_v^o, ω_v^o) to a latent variable, i.e. the face center. Δ_v and ω_v are respectively the offset and the corresponding weight of the vote. (Δ_v^o, ω_v^o) are similarly defined. The face center is localized by using the votes associated with all the landmarks (that is the votes from all image patches); this leads to a robust estimation of its location. Let us denote the estimated face center by $\hat{y}^o$ and assume a voting element v casts a vote at $y_v^o = y_v + \Delta_v^o$ with y_v the image location at which the voting element is extracted from, the sieving works as follows:

$$\omega_v := \omega_v \cdot \delta(\exp(-\frac{|y_v^o - \hat{y}^o|}{\beta}) > \lambda^o) \tag{1}$$

By the negative exponential function, we convert a distance measure in the range $[0, \inf)$ to a proximity measure in the range $(0, 1]$ with β a fixed parameter that controls the steepness of this function. λ^o is a threshold. Sieving can be interpreted as a filter that rejects the voting elements whose votes for the face center are far from the estimated center. The set associated with the landmark l and region r after the face center sieving is denoted by $\bar{V}_l$ and $\bar{V}^r$ respectively.

This procedure has been applied to effectively remove the *invalid* votes for facial feature detection. We adopt a similar idea in this work a) for estimating the predictive power of each segmented region as well as b) for estimating the reliability by which each of the facial landmarks is localized by the local-based RF.

2) RPP Estimation: It is difficult to pose the RPP estimation as a supervised classification problem as it is intractable to generate all types of occlusions. Here we take an unsupervised approach that estimates RPP from a set of features based on the region statistics and vote confidence. Specifically, we utilize the votes confidence calculated by the votes sieving procedure. Similarly to [38], we extract features directly from the voting maps as follows:

- $x_r^1 = \frac{\sum_{v\in\bar{V}^r}\omega_v}{\sum_{v\in V^r}\omega_v}$. This is the ratio of the sum of the vote weights in the segmented region r after and before the face center sieve is applied.
- $x_r^2 = U_r$. This is the area size of the region in pixels.
- $x_r^3 = \frac{U_r^{\text{box}}}{U_r}$. This is the fraction of the region that lies inside the face bounding box. U_r^{box} is the area of the region that lies inside the face bounding box. Roughly speaking, the smaller x_r^3 is, the more likely it is that r is an external object, i.e., an occluder of the face. In the example shown in Fig. 2, a large proportion of the hand region lies outside the bounding box, and therefore its RPP value is very low.

Given these features, we propose a rule-based method for calculating the RPP as follows. First, we identify the largest most likely face region. We do so by selecting the M largest regions inside the bounding box and assume that at least one of them belongs to the face. This is a reasonable assumption in real scenarios. From those M regions we select the one with the highest x_r^1 and put it in a set R_{SP}^0. We then put in R_{SP}^0 tiny regions, i.e. that satisfy $x_r^2 < \tau$ (where τ is to 50) and set the RPP of all regions in R_{SP}^0 to 1. The predictive

Algorithm 1 Cascaded Pose Regression (RCPR)

```
Input: Image $I$, initial pose $S^0$
Output: Estimated pose $S^T$
for $t$=1 to $T$ do
    $f^t = h^t(I, S^{t-1})$            ▷ Shaped-indexed features
    $\Delta S^t = R^t(f^t)$            ▷ Apply regressor $R^t$
    $S^t = S^{t-1} + \Delta S^t$       ▷ update pose
end for
```

power of all the other regions is estimated based on two strong cues: 1) the more inconsistent votes from one region, the lower RPP; 2) the bigger proportion of one region appears outside the face bounding box, the lower RPP. Formally, the RPP p_r of region r is defined as follows:

$$p_r = \begin{cases} 1 & \text{if } r \in R_{SP}^0 \\ \alpha + (1-\alpha) x_r^1 x_r^3 & \text{if } r \in R_{SP} \setminus R_{SP}^0 \end{cases}. \quad (2)$$

The product, $x_r^1 x_r^3$ is normalized to the range of [0, 1] in the set of $R_{SP} \setminus R_{SP}^0$ and is the main feature used for RPP estimation. The parameter α is the lower bound of the RPP, that is, the range of the RPP is [α 1]. We empirically set it to 0.2 and will discuss the sensitivity with respect to it in the experimental section.

B. Face Alignment Model Adaptation With RPP

In this section, we will first describe the original Cascaded Pose Regression (CPR) [8], [14] and Robust Cascaded Pose Regression (RCPR) [7] framework then we describe how the above RPP information is used to adapt these models in the presence of un-modeled occlusions.

1) CPR and RCPR Framework: The CPR framework has been shown to be effective and accurate in estimating the location of face landmarks [8], [14]. The procedure can be summarized as follows in Algorithm 1.

It starts from an initial shape S^0 and apply a sequence of regressors to update the shape until the last stage of regressor is applied. At the t-th iteration, the shape estimated at the previous iteration S^{t-1} is updated based on shape-indexed features $h^t(S^{t-1}, I)$, where I is the image. $S^t = S^{t-1} + \Delta S^t$ where ΔS^t is the shape update. As in [8], which is called Explicit Shape Regression (ESR), two-level cascaded regression is used, i.e., at each iteration, there are K primitive fern regressors $R^t = (R_1^t, ..., R_k^t, ..., R_K^t)$ that share the same input, namely features that are indexed relative to S^{t-1}, and whose outputs are combined in order to obtain the shape update ΔS^t as follows:

$$\Delta S^t = \sum_{k=1}^{K} \Delta S_k^t = \sum_{k=1}^{K} R_k^t(h^t(S^{t-1}, I)) \quad (3)$$

Robust Cascaded Pose Regression (RCPR) [7] improved CPR in three aspects: 1) it proposes a new interpolated shape-indexed feature, which is more robust to large shape variations; 2) it proposes a 'smart restart' scheme that deals with unreliable shape initializations; 3) it proposes area-based local regression to handle occlusion. Three typical variants are RCPR (feature only), RCPR (feature+restart) and RCPR (full). The area-based local regression (ferns) can be viewed as the third level of regression. It can be summarized as follows. Given the face location in an image, the face is divided into a 3 × 3 grid. Instead of training a single boosted regressor, N regressors are trained and each regressor is allowed to draw features only from 1 of the 9 pre-defined zones. Finally, each of the regressors proposed updates $\delta S_1, \cdots, \delta S_N$ are combined through a weighted mean voting. For the k-th update at the t-th iteration, the update of RCPR is calculated as:

$$\Delta S_k^t = \sum_{n=1}^{N} w_k^n \delta S_n^k. \quad (4)$$

where w_k^n is the weight that is inversely proportional to the occlusion estimation in the zones from which the regressor drew features.

2) Model Adaptation With RPP: Either the update is calcualted from Eq. 3 or Eq. 4, it is based on the shape-indexed features. There are k different ferns in Eq. 3 and N different ferns in Eq. 4. Note that despite the fact that the image features used by different weak regressors are indexed relative to the same pose, the weak regressors are different random ferns, and therefore the actual image features used by each regressor are at different pixel locations for each one. We first show how we use the RPP to adapt the update funtion of Eq. 3. Assuming F features are used by each fern regressor, we denote the image locations used to calculate the features of the k-th regressor as $\mathrm{x}^k = (\mathrm{x}_1^k, ..., \mathrm{x}_f^k, ..., \mathrm{x}_{2F}^k)$. In total, $2F$ pixel locations are used to produce F features. In Section III-A2 we have calculated the Regional Predictive Power, thus we can directly get the pixel predictive power according to which region it belongs to. The overall predictive power of the $2F$ locations is calculated as the mean value, that is

$$w_k = \frac{1}{2F} \sum_{f=1}^{2F} \sum_{r \in R} p_r \delta(\mathrm{x}_f^k \in r). \quad (5)$$

We adapt the regression model of Eq. 3 by reweighing the outputs of the K weak regressors by their respective predictive power. The above weight is normalized to $\bar{w}_k = \frac{K}{\sum_{k=1}^{K} w_k} w_k$, then the shape update at the t-th iteration is:

$$\Delta S^t = \sum_{k=1}^{K} \bar{w}_k \Delta S_k^t. \quad (6)$$

The first two variants of RCPR update their pose by Eq. 3 as well. Therefore our RPP-adapted version of RCPR (feature only) and RCPR (feature + restart) is adapted by the above equation.

The full version RCPR (full) uses Eq. 4 for pose update. We replace its weight w_k^n by our RPP-based weight $\bar{w}_k^n$. It is calculated in a similar way to Eq. 5. Then the update function is replaced by:

$$\Delta S_k^t = \sum_{n=1}^{N} \bar{w}_k^n \delta S_n^k. \quad (7)$$

To this end, we have shown how the RPP can be used to adapt both the two-level version of CPR (ESR) and three-level version of CPR (RCPR) and its variants.

C. Initialization From Local-Based Model

Existing iterative methods, e.g., the SDM [34] and CPR [14], depend on initialization and only those initializations that lie within a certain range can converge to the correct solution. However, there is no guarantee that the same face detector is used during the testing and training time. For instance, the SDM is trained based on mean pose deduced from Viola-Jones detector, however, Viola-Jones face detector misses many faces in the COFW dataset due to its heavy occlusion. Here we propose an initialization scheme that uses the estimated landmark locations and their estimated reliability, as those are provided by the local based Regression Forest method. Since the RF-based method is based on local patch features it does not require initialization, thus it is inherently more robust to face bounding box shifts.

Specifically, let us denote the estimate from the RF method in Section III-B by $y = (y_1, ..., , y_l, ..., y_L)$. Here, we also estimate the reliability of each landmark, that is, the confidence that the localization is correct. This differs from most of the face alignment methods. The reliability of a landmark is derived from the votes that are used to localize it and is calculated as follows:

$$s_l = \sum_{v \in \bar{V}_l} \omega_v \Big/ \sum_{v \in V_l} \omega_v \tag{8}$$

We then find the L_{com} common landmarks shared by the RF-based model and the RCPR model. Then instead of randomly selecting m shapes from the training set, we search the m nearest neighbors to the shape estimated by the RF. The distance between shapes is calculated as the sum of weighted Euclidean distances of all the common landmarks, where the weights are given by Eq. 8. This weighted distance measure suppresses the impact of the landmarks with large localization errors. Formally, the distance from the estimated shape vector y, to another shape y' is given by,

$$d(y, y') = \sum_{l=1}^{L_{com}} s_l ||y_l - y'_l||_2. \tag{9}$$

Note that, when calculating the distance, all the shapes are first normalized by procrustes analysis. This distance is used to calculate the m nearest neighbors in the training set - those are used to initialize the cascaded method. Conceptually this initialization scheme is similar to [24]. However, it uses backprojection to measure the similarity of a detection to the training samples for viewpoint estimation while our method measures the similarity in shape space and used it for selecting initialization shapes.

D. Method Summary

We summarize the proposed method in Algorithm 2. We emphasize that our method relies on two models, namely the regression forest **RF** and a model from CPR family, that is **Fern**, both of which do not need to be retrained. We only adapt the second model (i.e. Algorithm 1) using information derived from the first model in order to make it more robust to heavy occlusions. More specifically, first the regression forests are used to estimate the region predictive power (**RPP**) of local regions and to give good initializations S^0 of the shape. These are then used in an adapted **Fern** method, that is an adaptation of Algorithm 1. The adapted Fern method, starting from the initialization S^0, updates the pose using Eq. 6 (or Eq. 7) in step 3 of Algorithm 1. We note that our method outputs not only an accurate and robust face alignment but also a dense face mask that indicates which pixels belong to the face and which not.

Algorithm 2 The Proposed Framework

```
Input: Face image $I$, Face bounding box BB, Regression
    Forest RF, (R)CPR ferns Fern.
Output: Estimated pose $S^T$, Face Mask FM
  Do segmentation on $I$ and get superpixels $R_{sp}$
  RPP, $y \leftarrow$ RF($R_{sp}$, I, BB)        ▷ Get RPP and pose $y$
  Calculate init pose $S^0$ based on $y$
  $S^T \leftarrow$ Fern( $S^0$, RPP, I) ▷ Apply Fern as Algorithm 1,
    adapted by Eq. 6 or Eq. 7.
  FM $\leftarrow$ RPP                    ▷ Set threshold on RPP
```

IV. Experimental Results

A. Datasets and Implementation Details

We report the performance of our method on the most challenging datasets, namely, the Caltech Occluded Faces in the Wild (**COFW**) [7] dataset and the 300 Faces in-the-Wild (**300-W**) [27].

COFW is the most challenging dataset that is designed to depict faces in real-world conditions with partial occlusions [7]. The face images show large variations in shape and occlusions due to differences in pose, expression, hairstyle, use of accessories or interactions with other objects. All 1,007 images were annotated using the same 29 landmarks that are used for the LFPW [6] dataset. The training set includes 845 LFPW faces + 500 COFW faces, that is 1,345 images in total. The remaining 507 COFW faces are used for testing. Each image is annotated with the location of 29 facial landmarks and with corresponding 29 labels indicating whether the landmark is occluded or not. The average landmark occlusion on COFW is over 23%, while on LFPW is only 2%. Thus the occlusions in the test images are considerably more extended than in the training ones. We extend this dataset by providing the face masks for the 507 test images. The face mask indicates whether a pixel inside a face image belongs to the face (1) or not (0). Some example images are shown in Fig. 3.

300-W dataset is created for Automatic Facial Landmark Detection in-the-Wild Challenge [27]. Landmark locations for four popular data sets including LFPW, AFW, HELEN and XM2VTS, are re-annotated with the same 68 points mark-up. In addition, it contains a new set called iBug where

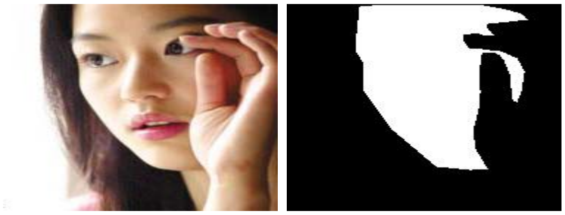


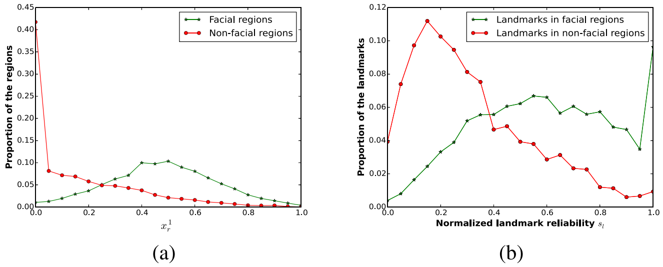
Fig. 3. Face image (left) and its mask annotation (right).

the images are more challenging. It provides a good benchmark for face alignment evaluation thus we make our comparison to the most recent methods based on this dataset. However, it only provides the training images for the challenge, thus we follow the experiment setting of [25] in order to compare with the recent methods. The training set is split into two parts. More specifically, the training part consists of AFW, the training images of LFPW and the training images of HELEN, with 3148 samples in total. The XM2VTS set is not used in our method as it is taken from very constrained environment and is not publicly available. The testing set consists of the test images of LFPW, the test images of HELEN and the images in the iBug set, with 689 samples in total. The test set is further partitioned into Easy-set (LFPW and HELEN test images) and Challenging-set (iBug images).

For the local Regression Forest, we use the trained model provided by [38], which is trained on a subset of AFLW [19] that contains mostly near frontal face images to ensure that the 19 facial landmarks are visible. We use all their default model parameters setting. Given that our adaptation methodology works on those models, it is clear that it does not exploit any training instances or annotations such as the occlusion labels. In our adaptation model, the number of the largest regions, that is the variable M in section III-A, is set to 3. The number of nearest neighbors that are used for initialization, that is, the variable m in section III-C is set to 5 - this is the default setting for RCPR. The error is measured as a fraction of the interocular distance. We note that in the evaluation process except when explicitly testing the face bounding box shift caused by changing the face detectors in Section IV-B5, the same face detector is used for both training and testing for fair comparison.

B. Results

1) RPP Estimation Evaluation: We empirically evaluate the performance of the RPP estimation based on the facial area annotation on COFW test images. Note that we do not use the annotation to tune our system during training. We set a threshold, equal to $\tau_{RPP} = \frac{1+\alpha}{2}$, on the RPP map. Regions with RPP value larger than the threshold are considered to be facial regions, and regions with smaller values are considered to be occlusions. Since we have annotated the face region masks for the testing images, we calculate the overlap area ratio inside the face bounding box to measure the performance, $\rho = \frac{A_{PPR} \cap A_{GT}}{A_{PPR} \cup A_{GT}}$. The average ratio is 72.4%, which is surprisingly high, given that the average percentage of area occlusion is 46.2%. We further infer the landmark occlusion state. If the RPP value of the region that one landmark is located is larger than a threshold τ_{RPP}, the landmark is regarded as visible, and vice versa. For landmark occlusion detection we get a 78/40% precision/recall, which is close to the 80/40% precision/recall reported in [7]. We note that in contrast to [7] we do not use occlusion information during training.



(a) (b)

Fig. 4. The distribution of x_r^1 feature (a) and landmark reliability s_l (b) for facial regions and non-facial regions. In (b) the value s_l of one face is normalized in the range between 0 and 1.

2) Feature Analysis: In Section III-A we developed features for RPP computing and reliability metric for landmark localization. We mainly rely on two features for RPP estimation, i.e. x_r^1 and x_r^3. In order to show the relevance of x_r^1, based on the face mask annotation, we plot the histogram of feature values for the face-regions and non-face regions, respectively, in Fig. 4(a). The p.d.f of x_r^1 in non-facial regions decreases gradually. On the contrary, the p.d.f of x_r^1 in facial regions peaks at around 0.5. In Fig. 4(b) we plot the histogram of the landmarks reliability s_l, defined in Eq. 8, from non-occluded and occluded face regions. We see that the reliability of most landmarks under occlusion tend to be lower than the reliability of the visible landmarks.

3) Face Alignment Evaluation on COFW: Here we evaluate the contribution of each component of the proposed method. We take four models from the CPR family as baseline methods: 1) the Explicit Shape Regression(ESR) [8]; 2) the feature only version of RCPR [7] (RCPR feature); 3) the RCPR with feature and smart restart [7] (RCPR feature+restart); 4) the full version of the RCPR (RCPR full). All of them are trained on the COFW training images with the same settings except the RCPR (full) which has used the landmark visibility labels during training. In the experimental comparison, **RF+baseline** is the direct combination of the RF sieving [38] and the baseline method, i.e. the output of [38] is used to find non-weighted nearest neighbouring shapes (all s_l in Eq. 8 are set to 1) to initialize the baseline methods. Their correponding model adaptation scheme is described in Section III-B2. **RPP weighted+RF initialization** is our full method. For methods not based on RF initialization we use 5 random initializations, that are the same for all methods. For the RF-based initialization methods, we replace the 5 initializations with the searched results. For the face images that need *smart restart*, the initializations in restart are all randomly generated. We repeat this process 4 times and report the average performance in terms of proportion of failures and average errors, similar to [7]. The number of restarts in the second round is also recorded as it is an important indicator of the efficiency. The results are shown in Fig. 5.

We can draw the following conclusions from the results: 1) the direct combination (**RF+baseline**) does not perform

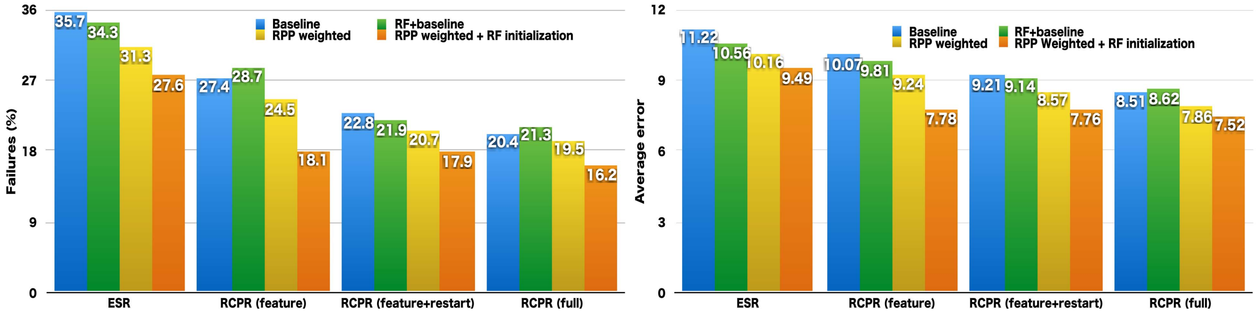


Fig. 5. Results on **COFW**, compared to CPR-family approaches [7], [8].

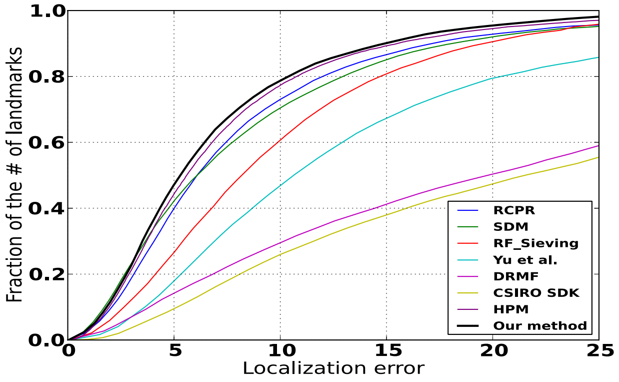


Fig. 6. Comparison to the recent methods, SDM [34], RCPR [7], RF_Sieving [38], method of Yu et al. [41], DRMF [3], and CSRIO SDK [11] and HPM [18] on COFW test images for their common 16 facial landmarks (only 15 for [38]). For the DRMF, the pre-computed face bounding box model is used since the tree-based method does not work on such images.

better than the baseline method; 2) the weighted models improve all the baseline methods in the CPR family, at an average mean error reduction of 0.8 and a decrease of failure rate of 2.6%; 3) it is worthy to note that the RPP based weights are even more effective than the original *learned* weights used in the RCPR (full) model, with a failure cases decrease 1% and a mean error decrease of 0.65; 4) the proposed initialization scheme is very effective and further decreases the mean error by 0.8 and the failure cases by 4%. 5) the smart restart has less impact when our proposed initialization scheme is applied. The number of restarts decreases from 200 to 30 among the 507 images, which means much fewer instances (85% less) require to restart the initializations [7]. The comparison to other state of the art methods on COFW is shown in Fig. 6, where the proposed method, that is built on top of RCPR (feature only), shows superior performance. Some examples are shown in Fig. 8.

We also compare to the recent methods that with codes publicly available on the common landmarks of the COFW test images, as shown in Fig. 6. For Hierarchical Deformable Part Model (HPM) [18] , since the code is not available, we communicate with the author and get the detection results, which is slightly better than what they have reported in the original paper. As can be seen, our proposed method shows competitive results on this challenging dataset.

In the proposed RPP model, there is one parameter α that influences the facial landmark localization. We increase its value from 0 to 1 with a step of 0.1 for the PCPR (feature+restart) model. The result is shown in Table I.

TABLE I

SENSITIVITY OF α

α	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Fail. rate	23.4	21.3	20.7	20.5	20.7	21.1	21.4	22.7	22.6	22.7	22.8

TABLE II

MEAN ERROR (68P) VS. # OF INITIALIZATIONS

m	1	2	3	4	5	6	7	8	9	10
Mean error	8.11	7.35	6.93	6.74	6.69	6.56	6.47	6.40	6.34	**6.25**

TABLE III

300-W DATASET (68 LANDMARKS)

Method	Full-set	Easy-set	Challenging-set
ESR[8]	7.58	5.28	17.00
SDM[34]	7.52	5.60	15.40
LBF fast[25]	7.37	5.38	15.50
LBF[25]	6.32	4.95	11.98
RCPR[7] (baseline)	7.54	5.67	15.50
Our method	6.69	5.50	11.57

TABLE IV

300-W DATASET (49 LANDMARKS)

Method	Full-set	Easy-set	Challenging-set
IFA[4]	7.48	5.58	15.30
SDM[34]	7.06	5.56	13.22
RCPR[7] (baseline)	7.20	5.47	14.28
Our method	6.57	5.40	11.40

When α is set to 0, the result is the worst, when α lies between 0.1 and 0.5, the performance is stable and when α becomes larger than 0.5, the performance approaches gradually to the baseline method, i.e., the model with equal weights. We set the value to 0.2 in all our experiment. Its value can be set by cross validation in practice.

4) Face Alignment Evaluation on 300-W Dataset: First on this dataset we evaluate the impact of the value of m in Section III-C. We vary the value of m from 1 to 10 and record the landmark-wise mean localization error of the 300W test images. As shown in Table II, the error decreases gradually with the increase of initialization number. In what follows, we set $m = 5$ since the baseline methods, both CPR and RCPR uses this value for runtime performance.

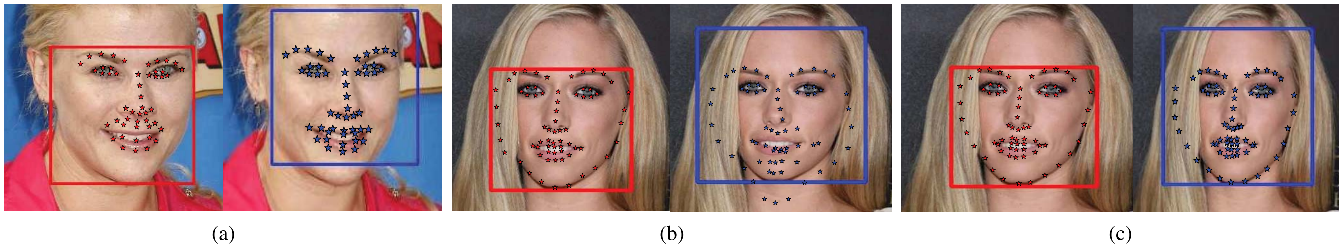


Fig. 7. Example results based on Viola-Jones face detector (blue) and 300-W face detector (red). SDM is trained based on Viola-Jones face detection and the other two are trained on 300-W face detection. The number under each pair shows increase of failure cases when face detection changes from one to the other. (a) SDM [34] (↑ 20%). (b) RCPR [7] (↑ 7%). (c) Our method (↑ 4%).

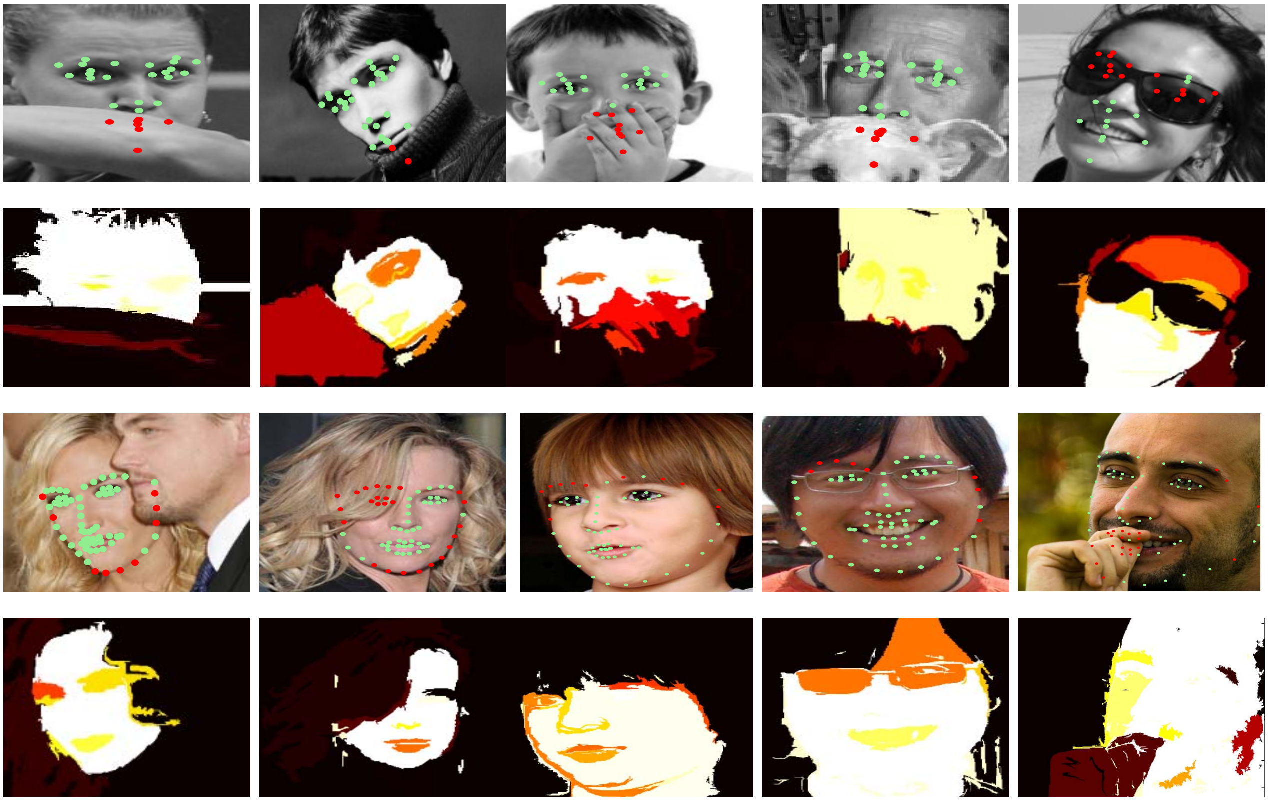


Fig. 8. Example results from COFW (first two rows) and LFPW and HELEN (last two rows), including landmarks detection results (upper) and the corresponding RPP map (lower). See Fig. 1 for color map definition.

We then compare our proposed method with the most competitive methods including the Supervised Descent Method (SDM) [34], the ESR[2] [8], the Incremental Face Alignment (IFA) [4] and the RCPR [7]. We use the full-RCPR version but we do not use any occlusion labels. For each of the regressor in Eq. 4, we treat them equally during the training stage and set the weight to $\frac{1}{N}$. This is equivalent to treat all landmarks visible. We take this as the baseline for adaptation as this gives us the best results compared to other RCPR variants. We first make the comparison as shown in Table III where the results of SDM, ESR, LBF and LBF-fast are quoted from [25]. We train the baseline RCPR model on the same training set for a fair comparison. As can be seen, although we only have comparable results to LBF, our results are better than the rest of the models. We note that LBF needs to train hundreds of thousands of trees. Taking this 68-landmark face as an example, its full model contains 5 stages and for each landmark, 1200 trees (with depth 7) are used. Thus in total, there are 1200×68×5 trees are needed, which is a huge number. On the contrary, in our method, both the local Regression Forest and the RCPR model is quite easy to train and the model size is much smaller. We also note that as shown in Table II, if we use more than 9 initializations, we obtain better performance than LBF in terms of localization accuracy (**6.25 vs. 6.32**). The improvement over the baseline RCPR model validates the effectiveness of our proposed method. We then compare to IFA and SDM in Table IV, as they show the state of the art results and have available test code. We train the baseline RCRP model on the Multi-PIE+LFPW (similar to the SDM model according the description of the paper) for localizing the 49 inner

[2] The result might be different from that in Section IV-B3, where the re-implementation source code is used.

facial landmarks. Then we apply our adaptation method on it. As can be seen, though the baseline RCPR method fails to compete the IFA and SDM, our method improves it clearly and shows better performance. From the two comparison we also note that the superior performance of our method on the Challenging subset is more significant, which is as expected since those images contain much more occlusions.

5) Face Bounding Box Shifts: Face detection itself is a challenging problem for faces under occlusion and with different head poses [21]. Thus for most of the methods we have discussed in this paper, face alignment starts from a given face bounding box. However, as different types of face detectors are available, there is no guarantee that the same detector is employed for both training and testing, we in this section evaluate the effect of face bounding box changes that is caused by different face detectors on the easy set of 300-W (LFPW and HELEN test images). As shown in Fig. 7, when the face bounding box changes, the performance of the cascaded methods changes significantly. This is as expected because the cascaded methods reply on face bounding box to calculate the initialization. The failure cases of the SDM method increases by 20% on average when the face bounding box of the test images changes from Viola-Jones face detector to 300-W face detector while that of the RCPR increases by 7% when the face bounding box changes the other way. The fact that the increase in failure of the SDM method is higher than that of the PCPR is probably due to their difference in initialization methodology, since the SDM only calculates one pose from the bounding box for initialization while the RCPR randomly selects 5 from the training instances. By using our proposed initialization scheme, the increase is minor (4%), around a half of the baseline RCPR method.

6) Run-Time: We record the run-time performance on a standard 3.30GHz CPU machine. For the COFW test images, the fps of the three components (segmentation (c++), Regression Forest (c++) and CPR (Matlab)) of our proposed method is 12, 17 and 11, respectively, and the overall speed is 4 FPS, that is a bit faster than the RCPR (full) method, and much faster than the HPM [18] (0.03FPS). On the LFPW and HELEN, the speed is 3.3 fps and 1 fps respectively while the segmentation takes longer time when the image becomes larger. Applying the segmentation only at a region of interest surrounding the face bounding box instead of the whole image can make our method more efficient. However, comparing to the LBF, which has reported 3000FPS execution time at testing stage, our method is still much slower. We will work towards improving the efficiency in our future work.

V. Conclusion

We present a method for face alignment model adaptation, based on Regional Predictive Power (RPP). We achieve the state of the art results for face alignment in challenging databases. Moreover, we show the efficacy of the proposed scheme in facial region prediction, something that can have applications in face analysis in real world applications such as face verification and facial expression recognition. In future work, we will integrate the face segmentation, facial region prediction and landmark estimation in a single optimization framework and extend the RPP for face analysis.

This work also raises a few interesting problems. First, with the rapid progress of face alignment, there is a demand of more advanced face detector that can work better in unconstrained environment, since most of the face alignment methods are based on face detection. Second, while most of the current methods work quite well on images with minor partial occlusion in a very fast speed but struggle under occlusion, developing a method based on the difficulty level of the test image to select a proper model is useful for practical applications.

References

[1] B. Amberg and T. Vetter, "Optimal landmark detection using shape models and branch and bound," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 455–462.

[2] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.

[3] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3444–3451.

[4] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1859–1866.

[5] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Continuous conditional neural fields for structured regression," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 593–608.

[6] P. N. Belhumeur, D. W. Jacobs, D. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 545–552.

[7] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1513–1520.

[8] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2887–2894.

[9] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[10] T. F. Cootes, M. C. Ionita, C. Lindner, and P. Sauer, "Robust and accurate shape model fitting using random forest regression voting," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 278–291.

[11] M. Cox, J. Nuevo-Chiquero, J. Saragih, and S. Lucey, "CSIRO face analysis SDK," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshop*, May 2013.

[12] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models," in *Proc. Brit. Mach. Vis. Conf.*, 2006, p. 6.

[13] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool, "Real-time facial feature detection using conditional regression forests," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2578–2585.

[14] P. Dollár, P. Welinder, and P. Perona, "Cascaded pose regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1078–1085.

[15] B. Efraty, C. Huang, S. K. Shah, and I. A. Kakadiaris, "Facial landmark detection in uncontrolled conditions," in *Proc. Int. Joint Conf. Biometrics*, Oct. 2011, pp. 1–8.

[16] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, 2004.

[17] H. K. Galoogahi, T. Sim, and S. Lucey, "Multi-channel correlation filters," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3072–3079.

[18] G. Ghiasi and C. C. Fowlkes, "Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1899–1906.

[19] M. Kostinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Nov. 2011, pp. 2144–2151.

[20] B. Martinez, M. F. Valstar, X. Binefa, and M. Pantic, "Local evidence aggregation for regression-based facial point detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1149–1163, May 2012.

[21] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 720–735.

[22] I. Patras and E. R. Hancock, "Coupled prediction classification for robust visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1553–1567, Sep. 2010.

[23] V. Rapp, T. Senechal, K. Bailly, and L. Prevost, "Multiple kernel learning SVM and statistical validation for facial landmark detection," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops*, Mar. 2011, pp. 265–271.

[24] N. Razavi, J. Gall, and L. Van Gool, "Backprojection revisited: Scalable multi-view object detection and similarity metrics for detections," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 620–633.

[25] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1685–1692.

[26] M.-C. Roh, T. Oguri, and T. Kanade, "Face alignment robust to occlusion," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops*, Mar. 2011, pp. 239–244.

[27] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 397–403.

[28] J. Saragih and R. Goecke, "A nonlinear discriminative approach to AAM fitting," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[29] B. M. Smith, J. Brandt, Z. Lin, and L. Zhang, "Nonparametric context modeling of local appearance for pose- and expression-robust facial landmark localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1741–1748.

[30] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3476–3483.

[31] P. A. Tresadern, P. Sauer, and T. F. Cootes, "Additive update predictors in active appearance models," in *Proc. Brit. Mach. Vis. Conf.*, 2010, p. 4.

[32] G. Tzimiropoulos and M. Pantic, "Optimization problems for fast AAM fitting in-the-wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 593–600.

[33] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial point detection using boosted regression and graph models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2729–2736.

[34] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 532–539.

[35] F. Yang, J. Huang, and D. Metaxas, "Sparse shape registration for occluded facial feature localization," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops*, Mar. 2011, pp. 272–277.

[36] H. Yang and I. Patras, "Face parts localization using structured-output regression forests," in *Proc. 11th Asian Conf. Comput. Vis.*, 2012, pp. 667–679.

[37] H. Yang and I. Patras, "Privileged information-based conditional regression forest for facial feature detection," in *Proc. 10th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Apr. 2013, pp. 1–6.

[38] H. Yang and I. Patras, "Sieving regression forest votes for facial feature detection in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1936–1943.

[39] H. Yang and I. Patras, "Fine-tuning regression forests votes for object alignment in the wild," *IEEE Trans. Image Process.*, vol. 24, no. 2, pp. 619–631, Feb. 2015.

[40] H. Yang, C. Zou, and I. Patras, "Face sketch landmarks localization in the wild," *IEEE Signal Process. Lett.*, vol. 21, no. 11, pp. 1321–1325, Nov. 2014.

[41] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas, "Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1944–1951.

[42] X. Yu, F. Yang, J. Huang, and D. N. Metaxas, "Explicit occlusion detection based deformable fitting for facial landmark localization," in *Proc. 10th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Apr. 2013, pp. 1–6.

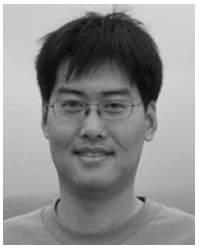
[43] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 94–108.

[44] X. Zhao, S. Shan, X. Chai, and X. Chen, "Cascaded shape space pruning for robust facial landmark detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1033–1040.

[45] F. Zhou, J. Brandt, and Z. Lin, "Exemplar-based graph matching for robust facial landmark localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1025–1032.

[46] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2879–2886.

Heng Yang (S'11) received the B.E. degree in simulation engineering and the M.Sc. degree in pattern recognition and intelligent system from the National University of Defense Technology, China, in 2009 and 2011, respectively. He is currently pursuing the Ph.D. degree with the School of Electronic Engineering and Computer Science, Queen Mary University of London, U.K. His research interests include pattern recognition, computer vision, and applied machine learning.



Xuming He (M'01) received the B.Sc. and M.Sc. degrees in electronics engineering from Shanghai Jiao Tong University, in 1998 and 2001, respectively, and the Ph.D. degree in computer science from the University of Toronto, in 2008. He held a post-doctoral position with the University of California at Los Angeles. He is currently a Senior Researcher with the Canberra Laboratory, National ICT Australia (NICTA), and an Adjunct Fellow with the Department of Engineering, Australian National University. His research interests include image segmentation and labeling, visual motion analysis, vision-based navigation, and undirected graphical models.

Xuhui Jia received the B.E. degree in software engineering from the Harbin Institute of Technology, China, in 2012. He is currently pursuing the Ph.D. degree with the Computer Vision and Machine Learning Group, The University of Hong Kong. His research interests include computer vision (face alignment, hand tracking) and applied machine learning (random forests, semi-supervised learning).

Ioannis Patras (SM'11) received the B.Sc. and M.Sc. degrees in computer science from the Department of Computer Science, University of Crete, Heraklion, Greece, in 1994 and 1997, respectively, and the Ph.D. degree from the Department of Electrical Engineering, Delft University of Technology, Delft, The Netherlands, in 2001. He is currently a Senior Lecturer with the School of Electronic Engineering and Computer Science, Queen Mary University of London, U.K. His current research interests are computer vision and pattern recognition, with an emphasis on the analysis of human motion, including the detection, tracking, and understanding of facial and body gestures and their applications in multimedia data management, multimodal human computer interaction, and visual communication. He is also an Associate Editor of the *Image and Vision Computing* journal.